

DOCUMENT RESUME *

ED 204 791

CS 206 496

AUTHOR Naccarato, Richard W.
TITLE A Generalizability Analysis of a Rhetoric Rating Experiment.
PUB DATE [72]
NOTE 12p.
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Evaluation Criteria: Higher Education: *Reliability: *Writing Evaluation: Writing Research

ABSTRACT

A study was initiated by a university rhetoric department to investigate the reliability of methods of rating student themes written as prerequisites for course exemptions. Ten experienced raters were selected to rate the beginning, middle, and ending paragraphs of 30 previously scored exemption themes that represented a range of achievement. Results of the study indicated that greater reliability of ratings could be obtained by using more raters to read selected paragraphs--specifically, ending paragraphs--than by using fewer raters scoring the entire themes. (AEA)

* Reproductions supplied by EDRS are the best that can be made
* from the original document.

ED204791

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)

* This document has been reproduced as
received from the person or organization
originating it.

☐ Minor changes have been made to improve
reproduction quality.

* Points of view or opinions stated in this docu-
ment do not necessarily represent official NIE
position or policy.

A GENERALIZABILITY ANALYSIS OF A RHETORIC RATING EXPERIMENT

Richard W. Naccarato
University of Iowa

"PERMISSION TO REPRODUCE THIS
MATERIAL HAS BEEN GRANTED BY

Richard W. Naccarato

TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC)."

CS206416

A Generalizability Analysis of a Rhetoric Rating Experiment
Richard W. Naccarato, University of Iowa

Like many freshman college programs throughout the nation, the Rhetoric Program at the University of Iowa provides arrangements for exempting students from certain basic courses if they are able to "pass" a theme exemption test. During the first week of classes in each semester, students may take a writing and a speech exam to exempt themselves from one or the other, or both, of these requirements. Those who pass these exams are given the appropriate amount of credit for each, and may not be required to take further coursework in the Rhetoric program. The subject of this experiment was the ratings of the writing portion of these exams, though the methods mentioned herein could apply as well to the ratings of the speech exams.

For the writing exam, students were given five general topics from which to choose and in a period of two hours they must select a topic, narrow it to a specific proposition that they can support within the time limits, and write an argumentative essay. These exams are then rated by experienced instructors, mostly teaching assistants, using a rating instrument that requires raters to focus on central idea and supporting material, organization, expression and mechanics. The raters are asked to combine their judgements of each of these categories and determine if the student passes or fails the writing exercise. Raters score their exams independently, then meet with a rating partner to reconcile scores where there is significant disagreement between the two raters scoring each essay. In some cases a third rater is called in to resolve differences of opinion between the original two raters. Needless to say, with hundreds of students writing exemption essays, this process is quite time consuming and yet must be carried out quickly so that students may be promptly notified of their exemption or reassignment status. This study was initiated by the Rhetoric Department to answer questions concerning the rating methods;

in particular, it will investigate the reliability of previous methods of making pass-fail judgements on the themes, and will propose alternative methods of rating based upon future generalizability decision studies.

METHOD

The investigators called for volunteers from experienced Rhetoric instructors with the invitation to take part in an experiment that would involve rating portions of student exemption themes. From these volunteers, ten raters were selected to rate portions of 30 student exemption themes. They were asked to work independently and to agree not to consult with anyone else about their responses. One of the questions which the Rhetoric Department wanted to investigate was whether the ratings of different portions of the themes would differ in pass-fail judgements. The raters were given three separate packets, one of beginning paragraphs (first paragraph of the theme), one of middle paragraphs (third paragraph), and one of ending paragraphs (last paragraph of the theme). The 30 student themes were selected randomly from low, middle and high-scoring strata, since they had previously been scored by two raters during the first week of classes. Raters were asked to make pass-fail judgements on one set of paragraphs at a time and, for each student, they were asked to write a word or phrase that indicated the "cue" which prompted their decision. A content analysis of these cues is currently being conducted but is not intended to be a part of this paper.

The raters were told that the thirty writing samples were drawn from exemption exams from the beginning of the semester and that they represented a range of achievement. Each sample paragraph was titled with the topic (from one of the five) that the student had chosen to write on. The directions to the raters included the following directive statement:

We might reasonably assume that an argumentative paper, more than papers in other modes, would have to be read completely and thoroughly in order to be rated analytically. We know, however, that the cues for success or lack of it come along quickly as we read into papers; a single rapid reading will often give us strong indication as to whether or not the papers will qualify the student as exempt...

Raters were then asked to think carefully about the criteria they were applying in their yes-no decision and to be specific in indentifying which cues lead them to their decision. Each rater then made ninety separate judgements (beginning, middle and ending paragraphs) and returned these ratings to the Rhetoric Department.

DESIGN

In generalizability terms, the unit of analysis for this study is topic - the rating of a student on any of the five topics upon which they wrote. The primary facets of interest were raters and students. A nested design was in effect since students were nested within topics, yet all 10 raters rated all three portions of the 30 student essays. The G-study design consisted then of raters crossed with students nested within essay topics. In such a design the rater by student variance is considered part of the residual variance; since all students were rated by the same raters, any interaction component between raters and students is considered error variance.

A limitation of this design was the lack of a student by topic interaction component, due to the nesting, and hence universe score variance could not be incremented by the interaction component to enhance dependability. Past research has found that requiring students to write on more than one topic tends to increase reliability of the ratings (Coffman, 1971).

Separate analyses of variance were computed for the ratings of beginning, middle and ending paragraphs. An analysis of variance was also computed for the total ratings of the essays, with "total rating" defined as the sum of separate ratings of beginning, middle and ending paragraphs. For these analyses all effects (facets) were assumed to be random, since it did not seem unreasonable that raters and students were representative of all possible raters and students; to assume a fixed or mixed effects model for future decision studies would not be difficult (Cronbach et al., 1972, ch. 4). Table 1 presents the results of these analyses of variance.

RESULTS AND DISCUSSION

Variance Components

A most obvious feature of Table 1 is the magnitude of the variance component estimates. The estimates are quite small since raters scored each paragraph on a pass-fail basis - a dichotomous rating score. The most serious consequence of such a scoring is the possible violation of normality assumptions, which leads to negative variance component estimates in some circumstances. Negative components were the result in this study for the Topic variable in the ratings of beginning and middle paragraphs. These components, though very slightly negative, were adjusted to zero according to Cronbach's recommendations (Cronbach et al, 1972, p. 57).

Another noticeable pattern in each of the analyses of variance tables are the differences in magnitude between the student and the rater variance component estimates. The ratio of these components ranges from about 7:2 for middle paragraph ratings to 10:1 for ratings of ending paragraphs. The ramification of these differences will be more evident as we look at the generalizability coefficients over students and raters.

Comparison of Generalizability Coefficients

Table 2 lists three generalizability coefficients of interest in this study. $\hat{e}\rho^2(R, S^*)$ is the coefficient of interest if we wish to generalize our results over different samples of raters, but assuming we have the same set of students each time that we rate in some future decision study. In most cases this coefficient would be of little interest, unless possibly we wanted to do repeated measures within any one semester. These values are listed in Table 2 primarily for comparative purposes with $\hat{e}\rho^2(R^*, S)$ and $\hat{e}\rho^2(R, S)$. The values of $\hat{e}\rho^2(R, S^*)$ reach an acceptable level (.50+) with 10 raters rating beginning paragraphs, but primarily with as few as two raters rating the ending paragraphs. The dependability of the ratings is also satisfactory with five raters rating all three paragraphs (as a total rating). The generalizability of ratings is considerably higher for $\hat{e}\rho^2(R, S^*)$ than for $\hat{e}\rho^2(R^*, S)$ or $\hat{e}\rho^2(R, S)$, however the assumption of a fixed set of students has limited practicability.

It is more reasonable that a department may use the same set of raters from one year to the next (at least for a few years) and, hence, $\hat{e}\rho^2(R^*, S)$ is not an unreasonable set of values to consider in a decision study. Assuming that we have the same set of raters rating a different set of students each semester, these data show that the dependability of ratings is greatest when rating middle or ending paragraphs. As for $\hat{e}\rho^2(R, S^*)$, the coefficients are largest when rating the last paragraphs of each theme. In fact, the summed ratings (total ratings) do not appear to be as reliable as the ratings of either the middle or ending paragraphs. Again, a satisfactory level of dependability is reached with as few as two raters rating either the middle or ending paragraphs.

The most general of the coefficients, $\hat{e}\rho^2(R, S)$, assumes that in a future decision study we would have a different set of raters rating a different set

of students. Obviously, the dependability coefficients for $\hat{\epsilon}_p^2(R,S)$ will be of smaller magnitude than either $\hat{\epsilon}_p^2(R,S^*)$ or $\hat{\epsilon}_p^2(R^*,S)$, since we are considering two random factors rather than just one. However, the values of $\hat{\epsilon}_p^2(R,S)$ are shockingly miniscule! $\hat{\epsilon}_p^2(R,S)$ never reaches an acceptable level unless we consider having five or more raters rating 20 or more students. Ratings of beginning and middle paragraphs have absolutely no dependability, and the summed ratings of all three paragraphs are hardly any better. It is suggested then, that if we wish to assume a new set of raters will be rating each new set of students, then only the ratings of ending paragraphs provide enough dependable information to make decisions.

CONCLUSIONS

Though an English Department may have difficulty in justifying the reading and scoring of only certain paragraphs within exemption themes, this study has shown that greater reliability of ratings may be obtained by employing more raters to read selected paragraphs within each theme than fewer raters scoring the entire themes. In particular, the greatest reliability is obtained when rating the ending paragraphs alone. The generalizability coefficients displayed herein for the ratings of ending paragraphs are quite comparable to interrater reliabilities for scoring essays found in other literature (Coffman, 1971). If we wish to assume that we would have a different set of raters rating a different set of students each semester, these data show little reliability of rating only certain paragraphs within an exemption theme - unless these are the ending paragraphs of the theme. If this is the case, this study implies that it would be more beneficial to have five raters (or more) score only the ending paragraphs than to use only a couple of raters to score entire themes. For some institutions with large exemption testing programs, such a rating procedure could be cost and

certainly time efficient. Before such a procedure could be implemented, however, it would be necessary to cross-validate these methods with current exemption procedures and other available criterion data.

References

Coffman, W.E. Essay Examinations, Educational Measurement, Edited by Robert L. Thorndike, American Council on Education, 1971, 271-302.

Cronbach, L.J., Gleser, G.C., Nanda, H., and Rajaratnam, N., The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles, New York: Wiley, 1972.

Table 1
Analyses of Variance Summary Tables

Beginning Paragraph Ratings					Middle Paragraph Ratings				
Source	SS	DF	MS	$\hat{\sigma}^2$	Source	SS	DF	MS	$\hat{\sigma}^2$
Topic	.585	3	.195	0	Topic	.295	3	.0983	0
Student (TOPIC)	14.752	116	.127	.0092	Student (TOPIC)	11.642	116	.1004	.0070
Rater	3.270	9	.363	.0023	Rater	4.070	9	.4522	.0025
Rater X	2.365	27	.088	.0018	Rater X	4.188	27	.1551	.0042
Topic Residual	36.265	1044	.035	.0350	Topic Residual	31.842	1044	.0305	.0305

Ending Paragraph Ratings					Total Paragraph Ratings [Beg+Mid+End]			
Source	SS	DF	MS	$\hat{\sigma}^2$	Source	SS	DF	$\hat{\sigma}^2$
Topic	1.617	3	.539	.0012	Topic	3.72	3	.0005
Student (TOPIC)	15.600	116	.134	.0106	Student (TOPIC)	94.70	116	.0681
Rater	2.217	9	.246	.0014	Rater	16.63	9	.016
Rater X	2.217	27	.082	.0018	Rater X	11.05	27	.0092
Topic Residual	29.267	1044	.028	.028	Topic Residual	138.852	1044	.1330

Table 2

Estimated Generalizability Coefficients for Various
Numbers of Raters and Students

BEGINNING PARAGRAPH RATING									
	$\hat{\epsilon}_p^2(R, S^*)$			$\hat{\epsilon}_p^2(R^*, S)$			$\hat{\epsilon}_p^2(R, S)$		
	$n'_s=10$	$n'_s=20$	$n'_s=30$	$n'_s=10$	$n'_s=20$	$n'_s=30$	$n'_s=10$	$n'_s=20$	$n'_s=30$
$n'_r=2$	.26	.21	.17	.25	.40	.50	.00	.00	.00
$n'_r=5$	.46	.39	.33	.18	.31	.40	.00	.00	.00
$n'_r=10$	.63	.56	.52	.12	.22	.30	.00	.00	.00
MIDDLE PARAGRAPH RATING									
	$\hat{\epsilon}_p^2(R, S^*)$			$\hat{\epsilon}_p^2(R^*, S)$			$\hat{\epsilon}_p^2(R, S)$		
	$n'_s=10$	$n'_s=20$	$n'_s=30$	$n'_s=10$	$n'_s=20$	$n'_s=30$	$n'_s=10$	$n'_s=20$	$n'_s=30$
$n'_r=2$	.16	.11	.08	.49	.65	.74	.00	.00	.00
$n'_r=5$	.33	.23	.18	.39	.56	.66	.00	.00	.00
$n'_r=10$	.49	.38	.29	.29	.46	.57	.00	.00	.00
ENDING PARAGRAPH RATING									
	$\hat{\epsilon}_p^2(R, S^*)$			$\hat{\epsilon}_p^2(R^*, S)$			$\hat{\epsilon}_p^2(R, S)$		
	$n'_s=10$	$n'_s=20$	$n'_s=30$	$n'_s=10$	$n'_s=20$	$n'_s=30$	$n'_s=10$	$n'_s=20$	$n'_s=30$
$n'_r=2$	.50	.52	.54	.46	.63	.71	.26	.36	.40
$n'_r=5$	.71	.73	.73	.49	.66	.71	.38	.51	.55
$n'_r=10$	.83	.82	.84	.51	.66	.73	.44	.57	.63
TOTAL RATING									
	$\hat{\epsilon}_p^2(R, S^*)$			$\hat{\epsilon}_p^2(R^*, S)$			$\hat{\epsilon}_p^2(R, S)$		
	$n'_s=10$	$n'_s=20$	$n'_s=30$	$n'_s=10$	$n'_s=20$	$n'_s=30$	$n'_s=10$	$n'_s=20$	$n'_s=30$
$n'_r=2$	.39	.33	.29	.28	.44	.54	.03	.05	.06
$n'_r=5$	.61	.55	.51	.21	.35	.44	.05	.09	.11
$n'_r=10$	.68	.67	.66	.16	.28	.38	.06	.11	.15